Jeannine Beeken
William Van Belle
Dirk Speelman
K.U.Leuven

Connectors : a Substantial Part of the CONST-Program

Abstract

The CONST-program is a computational writing tool for commercial and business communication in Dutch. It enables the user to consult and to integrate automated information during the processes of text-writing. The information is stored in two interactive modules : 1) a *tactical* component containing a) a CONST-internal lexicon containing information about 1000 connectors or discourse structuring items, b) an adapted part of the electronic version of the 'Van Dale' dictionary (1992), providing more (explanatory) information about the connectors, c) an adapted part of the electronic version of the 'connectors'-part of the Dutch Grammar *Algemene Nederlandse Spraakkunst* (1984) and d) a technical set providing information about the structuring of footnotes, quotations, etc. and 2) a *strategic* component generating stratified linguistic text structures and their corresponding tree diagrams. This strategic component provides e.g. a profound elaboration of the chosen writing strategy including a range of complex text-act verbs (e.g. 'describe, clarify') and a set of rhetorical relations (e.g. 'purpose, conclusion'). Through these relations a link can be established between the strategic and the tactical components, realised by means of the different types of connectors.

1. Introduction

The CONST-program enables the user to consult and to integrate automated information during the processes of text-structuring and text-formulation . The information is stored in two interactive modules, namely 1) the *strategic* module containing algorithmic procedures generating stratified text structures and their corresponding tree diagrams, and 2) the *tactical* module containing three types of lexicons, namely a) a CONST-specific text-linguistic lexicon, b) an adapted electronic version of the 'connectors'-part of the explanatory dictionary 'Van Dale' *(Groot Woordenboek der Nederlandse Taal,* 1992 (VD12)), c) an adapted electronic version of the 'connectors'-part of the syntactic lexicon or reference grammar *Algemene Nederlandse Spraakkunst* (ANS).

Briefly, the CONST-program is the realisation of the automated interaction between 1) an existing word-processor, 2) branching procedures by means of which a text structure and ipso facto a structured text can be built up, 3) the visualised representation of the chosen text structure, and 4) different types of lexicons providing information concerning the formulation of and functional cohesion within a structured text.

2. The strategic component

The hotspot–driven construction of a structured text is fairly stereotypical, in that choices are made constantly concerning, successively, type of text, the central theme or subtype, the purpose of the text, the characteristics of the addressees and the elaboration of the chosen writing strategy including various complex text act verbs and rhetorical relations. A tree diagram is built up simultaneously. This tree diagram is the visualised representation of the choices made concerning the different levels of text–structuring. The most interesting part of the CONST–program is the possibility to link certain (terminal) tree nodes of the strategic component with entries belonging to different types of lexicons, which provide a variety of useful information about the different types of text–cohesive items or connectors.

3. The tactical component

In addition to a CONST–specific lexicon, the tactical component offers an up–to–date electronic version of the explanatory dictionary of Dutch VD12 and an electronic version of the Dutch grammar ANS. The most important issues here are, on the one hand, the organisation of the CONST–specific lexicon which contains information about 1) the syntactical (sub) categorisation of the connectors, 2) the rhetorical relations (general as well as specific) that can be, text–linguistically seen, extensionalized by a connector, 3) the type of linking function (constituent, sentence, paragraph etc.) a connector has, and 4) the relevant and corresponding entry or entries in VD12 or ANS. On the other hand however, both the electronic lexicons, i.e. the explanatory dictionary and the grammatical dictionary, have been profoundly restructured, and filtered in such a way that only the essential and functional information is presented in a user–friendly way.

3.1 The CONST–dependent consultation of the lexicons

The bound or limited character of the rhetorical relations is due to the fact that both VD12 and ANS are opened through or from within the CONST–specific lexicon of connectors. In that way, based on the information gathered from the CONST–lexicon, only the relevant parts or passages of the VD12– and ANS–entries are selected and displayed. Consequently, the CONST–specific lexicon can be defined as, on the one hand, the starting link of the text formulation process as far as cohesion is concerned, and, on the other hand, as the terminal link of a specific text structure built up interactively and visualised by means of a tree diagram.

3.1.1 The CONST–specific lexicon

The CONST–specific lexicon contains a list of about 1000 connectors or text structuring items. Each connector is specified for

1) *headword* : e.g. 'omdat, naargelang, m.a.w., naar aanleiding van, ten tweede'
2) *key word(s)* : e.g. 'aanleiding, tweede' (or blank)
3) *syntactic category* : conjunction(al expression), adverb(ial expression), preposition(al expression), noun, adjective, verb(al expression), adverbial or structure–introductory clause, abbreviation
4) *rhetorical relation* (general and specific) : cause (reason, cause), consequence (result, effect), purpose (intention, aim, expectation), comparison (neutral, negative, positive, modification), contrast (adjustment, complement), concession (neutral, intensification), condition (neutral, negative, positive), restriction (specification, adjustment), extent (neutral, positive, negative, modification, superlative), manner, modality (neutral, positive, negative), time (moment, sequence, duration), place, conclusion, summary, specification (explanation, clarification, illustration, paraphrase), enumeration (alternative, gradation).
5) *type of connector* : link between a) constituents or sentences and/or b) paragraphs
6) *contrastive information* concerning conditions of usage and pragmatic conditions on valency

As mentioned earlier, the CONST–specific lexicon functions simultaneously as the terminal node of particular branches of the chosen text structure and as the starting node of the searching procedures in the electronic dictionary VD12 and grammar ANS. In that way, the CONST–specific lexicon is located on the abstract intersection of the strategic component (text–structuring) and the tactical component (text–formulation). Put differently, when the user has reached the lexicological modules from within the strategic component, the program automatically suggests which connectors are relevant in the given situation, based upon its knowledge gained from the different stages of the strategic component the user has been through. In this respect, the CONST–program and ipso facto the CONST– specific lexicon not only functions as an electronic consultation database, but also as a writing aid which makes suggestions concerning text structuring and text formulation, c.q. formal cohesion.

3.1.2 The electronic version of the explanatory dictionary 'Van Dale'

The result of a consultation of the VD12–dictionary is either a short summary or a detailed and exhaustive overview of the lexicographic information available. The information in question concerns 1) spelling (e.g. variants), 2) phonetics, 3) morphology (e.g. variants, compounds, plural, comparative, superlative, conjugation), 4) (lexical) semantics (word–level : e.g. identification, antonym, synonym, translation; example–level: e.g. identification, example, explanation; content/semantics: e.g. numeric classification and differentiation of syntactic homonyms and lexical–semantic polysemes; definition: e.g. metadefinition; internal and external references), 5) syntax (syntactic categorisation), 6) pragmatical conditions of usage and rules (labels for frequency, style, technical terminology etc.), 7) etymology, 8) encyclopaedic information and 9) lexicographic comment. The problem here was twofold. First, the VD12–labelling used to identify the type of information was too cryptically encoded to be displayed directly within a user–friendly windows–environment. Second, the (enormous) quantity of information had to be, on the one hand, cut down drastically in order to remain functional and, on the other hand, structured in a stratified way so that its comprehensibility is preserved.

Now, when consulting the explanatory VD12–dictionary from within the CONST–program, it is obvious that the program must have some keys at its disposal, in order to reduce the huge amount of information significantly. The most relevant keys are:

1. headword
2. key word(s)
3. syntactic category
4. rhetorical relation (both general and specific)
5. cohesive type

The most essential subcomponent of a partial sequential searching procedure can be summarised as follows:

1) if there are one or more *key words* available in the CONST–specific lexicon for a given entry, revalue each key word as a headword for the searching procedures in the VD12–dictionary; if not, the *headword* of the CONST–lexicon functions as headword for the searching procedures in the VD12–dictionary

2) list the different values of the label <synt> (which stands for syntactic category) and make an overview of the first numeric value (e.g. 4 for pronoun, 9 for conjunction)

   a) if the unique value of <synt> corresponds to the value of the field 'syntactic category' of the CONST–lexicon, then make a list of the

various numbers of contents (<betn>) and display the amount of values as short summaries of semantic descriptions
b) if the first numerical value occurs more than once, then find out what e.g. the rhetorical relation is and compare the result to the final numerical values of <synt> (e.g. 940 means subordinating conjunction of cause). Proceed as follows : scan the different types of definitions of the VD12–dictionary for words expressing the general and specific rhetorical relation in question (e.g. 'cause, causal, reason...' for cause). Finally, construct a list of the selected parts of a VD12–entry and display the information in a structured and stratified way. This means that level–1 information (syntactic category, semantic definitions, style or register) is displayed before level–2 information (examples, collocations, idioms, proverbs) and level–3 information (morphological–syntactical information, etymology, pronunciation, accent).

c) if there are different initial numeric values, make a choice based upon the syntactic value in the CONST–specific lexicon. After that, proceed as ordered in a) and b).

3.1.3 The electronic version of the Dutch grammar 'ANS'

As stated earlier, an adaptation and a restructuring of the ANS was unavoidable, in order to guarantee that the user wouldn't be burdened with redundant and non–functional information. In order to cope with the non–feasibility of the scanned ANS, three steps had to be taken. First, two different types of indexes had to be automated, namely a word index and an index containing conceptual expressions. Second, certain words had to be marked by means of a set of valued codes which correspond on a one–to–one basis to the codes used in the CONST–specific lexicon and the VD12–dictionary. Third, the information had to be reorganised and stratified in such a way that not only the contents of (sub)chapters and paragraphs has been paraphrased by means of key words, but also, and most importantly, the contents of each subparagraph and each part of a particular subparagraph (e.g. definition, examples, conditions on usage, exceptions to the rule, etc.). In concreto, this means that the contents of every (sub)chapter, (sub)paragraph and also every meaningful (sub)part has been translated into key words, which now function as entries for of the hotspot–driven selecting procedures.

3.2 The CONST–independent consultation of the lexicons

It is almost obvious that all three lexicons can be consulted independently of each other and of the strategic component. This means that every type of information can be looked at at any time. As far as the VD12–dictionary is concerned, two different searching procedures have been installed : 1) the *semasiological* procedure and 2) the *onomasiological* procedure. The input of the semasiological procedure is a specific headword (keyword), its output is a structured and stratified presentation of each substantial part of the information available (cf. supra). The input of the onomasiological procedure is either one parameter–value or a combination of parameter– values, e.g. definitions and descriptions, semantic categorisation criteria (is–a–part–of, is–a–kind–of, a specified rhetorical relation, etc.), register and style, word–length, syntactic category, etc. As far as ANS is concerned, a link has been realised with the CONST–lexicon and VD12. The parameters used for merging these three lexicons constitute a subpart of those used for the integration of VD12, namely 1) headword/keyword (by means of a word index), 2) syntactic category (to differentiate syntactic homonyms, by means of numeric encoding) and 3) rhetorical relation and text–structuring function (to differentiate semantic polysemes, by means of numeric encoding)

4. Conclusion

We have tried to give a brief overview of the two interacting components of the CONST–program, and especially of the stratified structure and function of three different types of lexicons, namely the CONST–specific lexicon of connectors, the electronic version of the explanatory Dutch dictionary 'Van Dale 12' and, to a smaller extent, the electronic version of the Dutch grammar *Algemene Nederlandse Spraakkunst.*

References

Amsler, R. 1984. "Machine readable Dictionaries". *Annual Review of Information Science and Technology.* 19:161–209.

Barrett, E. (ed.) 1988. *Text, ConText, and HyperText.* MIT Cambridge.

Beeken, J. 1992. "CONST : Computerondersteunde Schrijftechnieken". in N. Mars (ed), *Informatiewetenschap 1992.* Stinfon. Leiden.

Beeken, J., G. Geerts & W. Van Belle 1992. "The CONST–project : Computer Instructed Writing Techniques." in P. O'Brian Holt & N. Williams (eds.), *Computers and Writing : State of the Art.* Intellect & Kluwer Publishers, Oxford–Dordrecht.

Boguraev, B. & T. Briscoe (eds.) 1989. *Computational Lexicography for Natural Language Processing.* Longman, London–New York.

Daiute, C. 1985. *Writing and Computers.* Addison–Wesley, Massachussetts.

Dale R., E. Hovy, D. Rosner & O. Stock (eds.) 1992. *Aspects of Automated Natural Language Generation. Proceedings of the 6th International Workshop on Natural Language Generation.* Berlin.

Dodd, S. 1988. "The Exeter Coditext Project" *Lexicographica* 4:11–18

Geerts, G., W. Haeseryn, J. de Rooij & M.C. van den Toorn 1984. *Algemene Nederlandse Spraakkunst.* Wolters–Noordhoff, Groningen–Leuven.

Haiman, J. & S. Thompson 1988. *Clause Combining in Grammar and Discourse.* Benjamins, Amsterdam.

Hess, K., J. Brustkern & W. Lenders. 1983. *Maschinenlesbare deutsche Wörterbücher. Dokumentation, Vergleich, Integration.* Niemeyer, Tübingen.

Heyn, M. 1992. *Zur Wiederverwendung maschinenlesbarer Wörterbücher.* Lexicographica Series Maior 45. Niemeyer, Tübingen.

Hovy, E. 1988. *Generating Natural Language under Pragmatic Constraints.* Hillsdale.

McKeown, K. 1985. *Text Generation. Using Discourse Strategies and Focus Constraints to Generate Natural Language Text.* Cambridge.

Lamers, H. 1989. *Handleiding voor beleidsteksten. Een handleiding om beleidsnota's, beleidsbrieven, jaarverslagen en notulen op te stellen.* Muiderberg.

McNaught, J. 1988. "Computational Lexicography and Computational Linguistics." *Lexicographica* 4:19–33.

Meteer, M. 1991. "Bridging the Generation Gap between Text Planning and Linguistic Realization." *Computational Intelligence* 7:296–304.

Nederhoed, P. 1984. *Helder Rapporteren. Een handleiding voor het schrijven van rapporten, scripties, nota's en artikelen in wetenschap en techniek.* Deventer.

Paris, C., W. Swartout & W. Mann (eds.) 1991. *Natural Language Generation in Artificial Intelligence and Computational Linguistics.* Dordrecht.

Phillips; M. 1985. *Aspects of Text Structure. An Investigation of the Lexical Organization.* North–Holland, Amsterdam.

Van Dale Groot woordenboek der Nederlandse Taal 1992. Utrecht.

Vanderveken, D. 1990. *Meaning and Speech Acts. Volume I : Principles of Language Use.* Cambridge.

Vanderveken, D. 1991. *Meaning and Speech Acts. Volume II : Formal Semantics of Success and Satisfaction.* Cambridge.

Wiegand, H. 1990. "The Dictionary as Text.' *Lexicographica* 6:1–126.

Zock, M. & G. Sabah (eds.) 1988. *Advances in Natural Language Generation.* Printer, London.